PRD: Analysis of very large data sets from relativistic heavy ion collisions

Key emerging challenges

- Routine analysis of large PB-scale analysis-ready datasets: e.g. ~10^9 x 1MB events
  - growth in complex analyses that exploit rich event structure
  - still require reasonable turn around times
- Harvesting new processing capabilities from heterogeneous architectures

Summary of research direction

- Continue to leverage event-level parallelism with distributed processing model
- Current datasets are ~10-100x smaller, then
  - filtered by investigation
  - duplicated & distributed by investigation
- How to easily track & augment event info from independent investigations?

Potential impact on software/systems

- low-overhead event-level data cataloging
- Expand use of:
  - Distributed file systems
  - Distributed/parallel analysis facilities
- Common algorithms/data-structures/ processing methods mapped to heterogeneous architectures

Potential impact on science communities or DOE capabilities

- Increased scientific productivity
- new analyses beyond current processing capabilities would become viable